CASIA-SURF: A Dataset and Benchmark for Large-scale Multi-modal Face Anti-spoofing

Shifeng Zhang[1], Xiaobo Wang[2], Ajian Liu[3], Chenxu Zhao[2], Jun Wan[1]*,
Sergio Escalera[4], Hailin Shi[2], Zezheng Wang[5], Stan Z. Li[1]
[1]CBSR, NLPR, CASIA, China; [2]JD AI Research; [3]M.U.S.T, Macau, China; [4]CVC, UB, Spain; [5]JD Finance.
{shifeng.zhang,jun.wan,szli}@nlpr.ia.ac.cn

Abstract

*Face anti-spoofing is essential to prevent face recognition systems from a security breach. Much of the progresses have been made by the availability of face anti-spoofing benchmark datasets in recent years. However, existing face anti-spoofing benchmarks have limited number of subjects (≤ 170) and modalities (≤ 2), which hinder the further development of the academic community. To facilitate future face anti-spoofing research, we introduce a large-scale multi-modal dataset, namely CASIA-SURF, which is the largest publicly available dataset for face anti-spoofing both in terms of subjects and visual modalities. Specifically, it consists of $1,000$ subjects with $21,000$ videos and each sample has 3 modalities (*i.e.*, RGB, Depth and IR). Associated with this dataset, we also provide concrete measurement set, evaluation protocol and training/validation/testing subsets, developing a new benchmark for face anti-spoofing. Moreover, we present a new multi-modal fusion method as a strong baseline, which performs feature re-weighting to select the more informative channel features while suppressing less useful ones for each modal. Extensive experiments have been conducted on the proposed dataset to verify its significance and generalization capability. Dataset is available at https://sites.google.com/qq.com/chalearnfacespoofingattackdete/.*

1. Introduction

Face anti-spoofing is an important problem in computer vision, which aims to determine whether the captured face is a real or fake face in the face recognition system. With the development of deep convolutional neural network (CNN), face recognition [33, 2] has achieved near-perfect recognition performance and already been applied in our daily life, such as phone unlock, access control, face payment, etc. However, these face recognition systems are very easy to be attacked in various ways including print attack, video replay attack, and 2D/3D mask attack, which cause the result of the recognition to become unreliable. Therefore, face presentation attack detection (PAD) [45, 3, 4] is a vital step to ensure that face recognition systems are in a safe reliable condition prior to face verification.

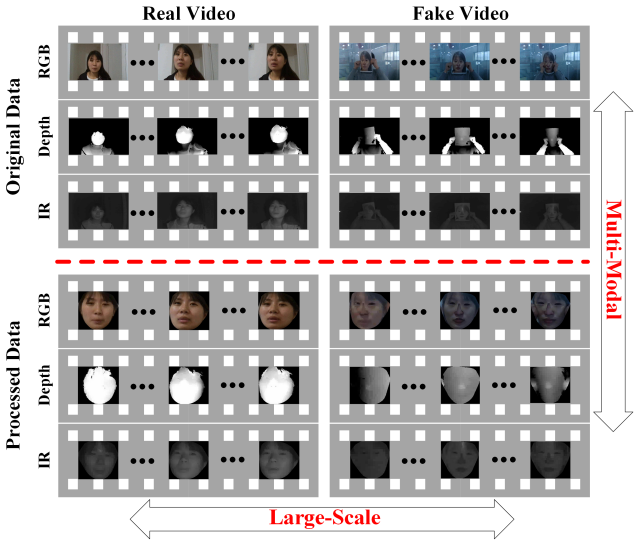

Figure 1. The CASIA-SURF dataset. It is a large-scale and multi-modal dataset for face anti-spoofing, consisting of $492,522$ images with 3 modalities (*i.e.*, RGB, Depth and IR).

Recently, face PAD algorithms [19, 31] have achieved great performances. One of the key point to this success is these existing datasets, which have contributed to spur interest and progress in face anti-spoofing research. However, compared with image classification [13] and face recognition [48] datasets including large-scale subjects and images, the public face anti-spoofing datasets are less than 170 subjects and $60,000$ video clips, as shown in Table 1. The limited number of subjects has greatly decreased the generalization ability and is far away from satisfaction for the requirements of practical applications. Besides, from Table 1, another problem is the limited data modality. Most

*Corresponding author

Dataset	Year	# of subjects	# of videos	Camera	Modal types	Spoof attacks
Replay-Attack [6]	2012	50	1200	VIS	RGB	Print, 2 Replay
CASIA-MFSD [49]	2012	50	600	VIS	RGB	Print, Replay
3DMAD [14]	2013	17	255	VIS/Kinect	RGB/Depth	3D Mask
MSU-MFSD [45]	2015	35	440	Phone/Laptop	RGB	Print, 2 Replay
Replay-Mobile [9]	2016	40	1030	VIS	RGB	Print, Replay
Msspoof [8]	2016	21	4704*	VIS/NIR	RGB/IR	Print
Oulu-NPU [5]	2017	55	5940	VIS	RGB	2 Print, 2 Replay
SiW [31]	2018	165	4620	VIS	RGB	2 Print, 4 Replay
CASIA-SURF (Ours)	2018	**1000**	**21000**	RealSense	**RGB/Depth/IR**	Print, Cut

Table 1. The comparison of the public face anti-spoofing datasets ($*$ indicates this dataset only contains images, not video clips).

of the current datasets only have one modal (*e.g.*, RGB), and the existing available multi-modal datasets [14, 8] are scarce with no more than 21 subjects. That is also a big problem hindering novel technology developments.

To solve these drawbacks, we introduce a large-scale multi-modal face anti-spoofing dataset, namely CASIA-SURF, which consists of $1,000$ subjects and $21,000$ video clips with 3 modalities (RGB, Depth, IR). It has 6 types of photo attack combined by multiple operations, *e.g.*, cropping, bending the print paper and stand-off distance. Some samples are shown in Fig. 1. As shown in Table 1, our dataset has two advantages: (1) It is the largest one in term of number of subjects and videos; (2) Our dataset has three modalities (*i.e.*, RGB, Depth and IR).

Another concern problem is how to judge the performance in face anti-spoofing. Many works [31, 19, 5, 9] adopt the Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER) and Average Classification Error Rate (ACER) as the evaluation metric, in which APCER and BPECER are used to measure the error rate of fake or live samples, respectively. While ACER is the average value of APCER and BPCER. However, in real applications, one may be more concerned about the false positive rate, *i.e.*, attacker is treated as real/live one. Inspired by face recognition [30], the Receiver Operating Characteristic (ROC) curve is introduced for large-scale face anti-spoofing in our dataset, which can be used to select a suitable threshold to trade off the false positive rate (FPR) and true positive rate (TPR) according to the requirement of real applications.

To sum up, the contributions of this paper are summarized as follows: (1) We present a large-scale multi-model dataset for face anti-spoofing. It has $1,000$ subjects that is larger at least 6 times than the public datasets, with three modalities (RGB, Depth and IR). (2) We introduce a new multi-modal fusion method to effectively merge the involved 3 modalities, which performs modal-dependent feature re-weighting to select the more informative channel features while suppressing less useful ones for each modality. (3) We conduct extensive experiments on the proposed CASIA-SURF dataset to verfiy its significance and generalization capability.

2. Related Work

2.1. Datasets

Most of existing face anti-spoofing datasets only have the RGB modalitiy including Replay-Attack [6] and CASIA-FASD [49], which are two widely used PAD datasets. Even the recently released SiW [31] dataset is no exception, which is collected with high resolution and image quality. With the widespread application of face recognition in mobile phones, there are also some RGB datasets recorded by replaying face video with smartphone or laptop, such as MSU-MFSD [45], Replay-Mobile [9] and OULU-NPU [5].

As attack techniques are constantly upgraded, some new types of presentation attacks (PAs) have emerged including 3D [14] and silicone masks [2], which are more realistic than traditional 2D attacks. Therefore, the drawbacks of visible cameras are revealed when facing these realistic face masks. Fortunately, some new sensors have been introduced to provide more possibilities for face PAD methods, such as depth cameras, muti-spectral cameras and infrared light cameras. Kim *et al.* [22] aim to distinguish between the facial skin and mask materials by exploiting their reflectance. Kose *et al.* [27] propose a 2D+3D face mask attacks dataset to study the effects of mask attacks, but the realted dataset is not public. 3DMAD [14] is the first publicly available 3D masks dataset, which is recorded using Microsoft Kinect sensor and consists of the Depth and RGB modalities. Another multi-modal face PAD dataset is Msspoof [8] that contains visible (VIS) and near-infrared (NIR) images of real accesses and printed spoofing attacks with ≤ 21 objects.

However, these existing datasets in the face PAD community have two common limitations as follows. Firstly, they all have the limited number of subjects and samples, resulting in the potential over-fitting risk when face PAD algorithms are tested on these datasets [6, 49]. Secondly, most of the existing datasets are captured by visible camera that only includes the RGB modality, causing a substantial portion of 2D PAD methods to fail when facing new types of PAs (3D and custom-made silicone masks).

2.2. Methods

Face anti-spoofing has been studied for decades. Some previous work [35, 42, 24, 1] attempt to detect the evidence of liveness (*i.e.*, eye-blinking). Another works are based on contextual [36, 25] and [43, 12, 21] information. To improve the robustness to illumination variation, some algorithms adopt HSV and YCbCr color space [3, 4], and Fourier spectrum [28]. All fo these methods use handcrafted features, such as LBP [34, 7, 47, 32], HoG [47, 32, 39] and GLCM [39]. They are fast enough and have relatively satisfactory performance only on small public face spoof datasets with poor generalizability.

Some fusion methods are propposed to obtain a more general countermeasure effective against a variation of attack types. Tronci *et al*. [41] propose a linear fusion at a frame and video level combination between static and video analysis. Schwartz *et al*. [39] introduce feature level fusion by using Partial Least Squares (PLS) regression based on a set of low-level feature descriptors. Some other works [10, 26] obtain an effective fusion scheme by measuring the level of independence of two anti-counterfeiting systems. However, these fusion methods focus on score or feature level, not modality level, due to the lack of multi-modal datasets. It is urgent to propose a multi-modal dataset.

Recently, CNN-based methods [15, 29, 37, 46, 31, 19] are presented in face PAD community. They treat face PAD as a binary classification problem and achieve remarkable improvements in the intra-testing. Liu *et al*. [31] design a novel network architecture to leverage two auxiliary information (the Depth map and rPPG signal) as supervision with the goals of improved generalization. Amin *et al*. [19] introduce a new perspective for solving the face anti-spoofing by inversely decomposing a spoof face into the live face and the spoof noise pattern. However, they exhibit a poor generalization ability during the cross-testing due to the over-fitting to training data. This problem has still not been solved, even if some works [29, 37] adopt transfer learning to train the CNN model from ImageNet [13]. These works bring us the insight that we need to collect a larger PAD dataset.

3. CASIA-SURF Dataset

As aforementioned, all of datasets involve fewer subjects and most of them contain one modality. Although the publicly available datasets have driven the development of face PAD and continue to be valuable tools for this community, it still leads to severely impede the development of face PAD. Especially, it still remains challenging problems under some conditions that require higher recognition accuracy, *e.g.*, face payment or unlock.

In order to address current limitations in PAD, we collect a new face PAD dataset, namely the CASIA-SURF dataset.

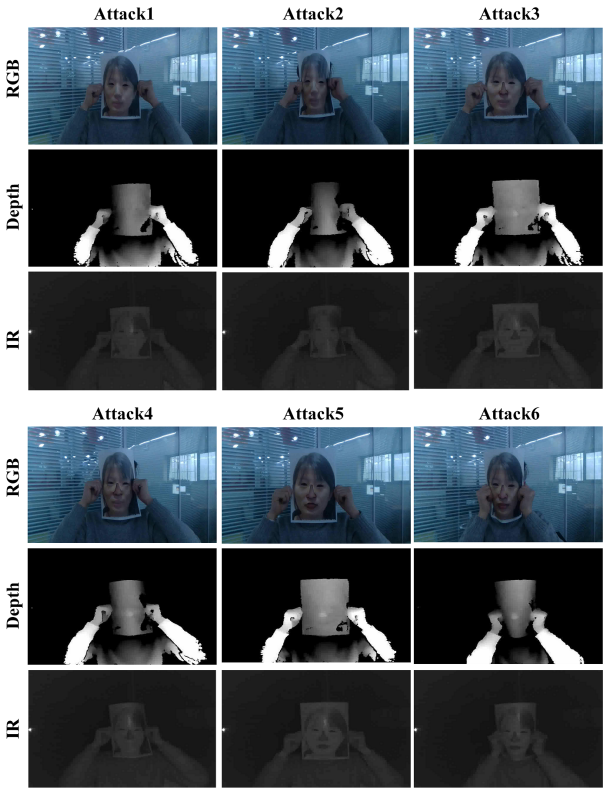

Figure 2. Six attack styles in the CASIA-SURF dataset.

To the best our knowledge, CASIA-SURF dataset is currently the largest face anti-spoofing dataset, which contains $1,000$ Chinese people in $21,000$ videos. Another motivation in creating this dataset, beyond pushing the research on face anti-spoofing, is to explore recent face spoofing detection models performance when considering with large-scale data. In the proposed dataset, each sample includes 1 live video clip, and 6 fake video clips under different attack ways (one attack way per fake video clip). In the different attack styles, the printed flat or curved face images will be cut eyes, nose, mouth areas, or their combinations. Finally, 6 attacks are generated in the CASIA-SURF dataset. Fake samples are shown in Fig. 2. Detailed information of the 6 attacks is given below.

- Attack 1: One person hold his/her flat face photo where eye regions are cut from the printed face.
- Attack 2: One person hold his/her curved face photo where eye regions are cut from the printed face.
- Attack 3: One person hold his/her flat face photo where eyes and nose regions are cut from the printed face.
- Attack 4: One person hold his/her curved face photo where eyes and nose regions are cut from the printed face.

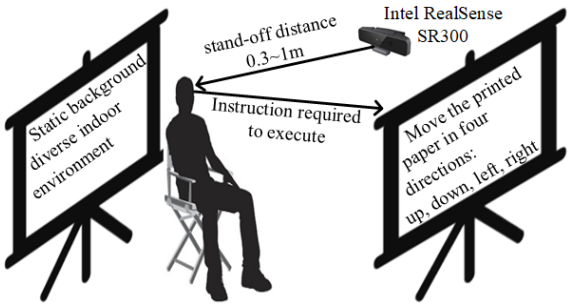

Figure 3. Illustrative sketch of recordings setups in the CASIA-SURF dataset.

- Attack 5: One person hold his/her flat face photo where eyes, nose and mouth regions are cut from the printed face.
- Attack 6: One person hold his/her curved face photo where eyes, nose and mouth regions are cut from the printed face.

3.1. Acquisition Details

We use the Intel RealSense SR300 camera to capture the RGB, Depth and Infrared (IR) videos simultaneously. In order to obtain the attack faces, we print the color pictures of the collectors with A4 paper. During the video recording, the collectors are required to do some actions, such as turn left or right, move up or down, walk in or away from the camera. Moreover, the face angle of performers are asked to be less 30^0. The performers stand within the range of 0.3 to 1.0 meter from the camera. The diagram of data acquisition procedure is shown in Fig. 3, where it shows how to record the multi-modal data via Intel RealSence SR300 camera.

Four video streams including RGB, Depth and IR images are captured at the same time, plus the RGB-Depth-IR aligned images using RealSense SDK. RGB, Depth, IR and aligned image are shown in the first column of Fig. 4. The resolution is 1280×720 for RGB images, and 640×480 for Depth, IR and aligned images.

3.2. Data Preprocessing

In order to make the dataset more challenging, we remove the complex background except face areas from original videos. Concretely, as shown in Fig. 4, the accurate face area is obtained through the following steps. Although there is lack of face detection for Depth and IR face images, we have a RGB-Depth-IR aligned video clip for each sample. Therefore, we first use Dlib [23] to detect face for every frame of RGB and RGB-Depth-IR aligned videos, respectively. The detected RGB and aligned faces are shown in the second column of Fig. 4. After face detection, we apply the PRNet [16] algorithm to perform 3D reconstruction and density alignment on the detected faces. The accurate face area (namely, face reconstruction area) is shown in the third column of Fig. 4. Then, we define a binary mask based on non-active face reconstruction area from previous step. The binary masks of RGB and RGB-Depth-IR images are shown in the fourth column of Fig. 4. Finally, we obtain face area of RGB image via pointwise product between RGB image and RGB binary mask. The Depth (or IR) area can be calculated via the pointwise product between Depth (or IR) image and RGB-Depth-IR binary mask. The face images of three modalities (RGB, Depth, IR) are shown in the last column of Fig. 4.

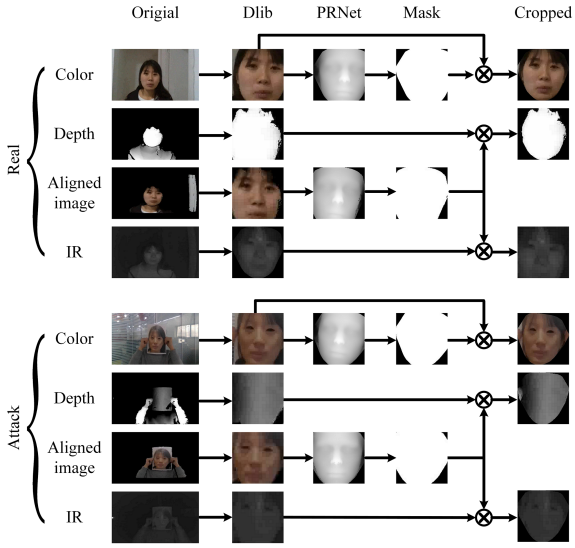

Figure 4. Preprocessing details of the three modalities of the CASIA-SURF dataset.

3.3. Statistics

Table 2 presents the main statistics of the CASIA-SURF dataset:

(1) There are $1,000$ subjects and each one has a live video clip and six fake video clips. Data contains variability in terms of gender, age, glasses/no glasses, and indoor environments.

(2) Data is split in three sets: training, validation and testing. The training, validation and testing sets have 300, 100 and 600 subjects, respectively. Therefore, we can get $6,300$ ($2,100$ per modality), $2,100$ (700 per modality), $12,600$ ($4,200$ per modality) videos for its corresponding set.

	Training	Validation	Testing	Total
#Obj.	300	100	600	1000
#Videos	6,300	2,100	12,600	21000
#Ori. img.	1,563,919	501,886	3,109,985	5,175,790
#Samp. img.	151,635	49,770	302,559	503,964
#Crop. img.	148,089	48,789	295,644	492522

Table 2. Statistical information of the proposed CASIA-SURF dataset.

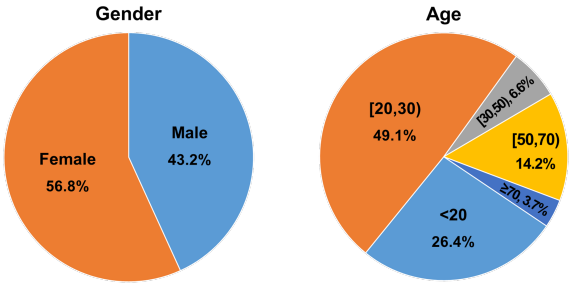

Figure 5. Statistical gender and age distribution of the CASIA-SURF dataset.

(3) From original videos, there are about 1.5 million, 0.5 million, 3.1 million frames in total for training, validation, and testing sets, respectively. Owing to the huge amount of data, we select one frame out of every 10 frames and formed the sampled set with about $151K$, $49K$, and $302K$ for training, validation and testing sets, respectively.

(4) After data prepossessing in Sec. 3.2, by removing non-detected face poses with extreme lighting conditions, we finally get about $148K$, $48K$, $295K$ frames for training, validation and testing sets on the CASIA-SURF dataset, respectively.

All subjects are Chinese peoples, and the information of gender statistics is shown in the left side of Fig.5. It shows that the ratio of female is 56.8% while the ratio of male is 43.2%. In addition, we also show age distribution of the CASIA-SURF dataset in the right side of Fig 5. One can see a wide distribution of age ranges from 20 to more than 70 years old, while most of subjects are under 70 years old. On average, the range of $[20, 30)$ ages is dominant, being about 50% of all the subjects.

3.4. Evaluation Protocol

Intra-testing. For the intra-testing protocol, the live faces and Attacks 4, 5, 6 as the final training set used to train the algorithm models. Then, the live faces and Attacks 1, 2, 3 are used as the validation and testing sets. The validation set is used for model selection and the testing set for final evaluation. This protocol is used for the evaluation of face anti-spoofing methods under controlled conditions, where training and testing set belong to the CASIA-SURF dataset. The main reason that we select the different attack types in the training and testing set is to increase the difficulty of face anti-spoofing detection task. In our experiment section, we will prove that there is still a big space to improve the performance under the ROC evaluation metric, especially, how to improve the true positive rate (TPR) at the little value of false positive rate (FAR), such as FAR=10^{-5}.

Cross-testing. The cross-testing protocol uses the training set of CASIA-SURF to train the deep models, which are then fine-tuned on the target training dataset (*e.g.*, the training set of SiW [31]). Finally, we test the fine-tuned model on the target testing set (*e.g.*, the testing set of SiW [31]). The cross-testing protocol aims at simulating performance in real application scenarios involving high variabilities in appearance and having a limited number of samples to train the model.

4. Proposed Approach

Before delving into our new dataset, we first build a strong baseline method as a tool for our experiment analyses. We aim at finding a straightforward architecture that provides good performance in our CASIA-SURF dataset. Thus, we regard the face anti-spoofing problem as a binary classification task (fake *v.s* real) and conduct the experiments based on the ResNet-18 [17] classification network. The ResNet-18 network consists of five convolutional blocks (namely res1, res2, res3, res4, res5), a global average pooling layer and a softmax layer, which is a relatively shallow network but has strong classification capabilities.

4.1. Naive Halfway Fusion

As described before, CASIA-SURF is characterized by multi-model (*i.e.*, RGB, Depth, IR) and the main research point is how to fuse the complementary information between these three modalities. We use a multi-stream architecture with three subnetworks to study our dataset with three modalities, in which RGB, Depth and IR data are learnt separately by each stream, and then shared layers are appended at a point to learn joint representations and cooperated decisions. The halfway fusion is one of the commonly used fusion methods, which combines the subnetworks of different modalities at a later stage, *i.e.*, immediately after the third convolutional block (res3) via the feature map concatenation. In this way, features from different modalities can be fused to perform classification. However, direct concatenating these features cannot make full use of the characteristics between different modalities.

4.2. Squeeze and Excitation Fusion

Since different modalities have different characteristics: the RGB data have rich details, the Depth data is sensitive to the distance between the image plane and the corresponding face, and the IR data measures the amount of heat radiated from a face. These three modalities have different advantages and disadvantages for different ways of attack. Inspired by [18], we propose the squeeze and excitation fusion method that uses the "Squeeze-and-Excitation" branch to enhance the representational ability of the different modalities' feature by explicitly modelling the interdependencies between their convolutional channels.

As shown in Figure 6, our squeeze and excitation fusion method has a three-stream architecture and each subnetwork is feed with the image of different modalities. The

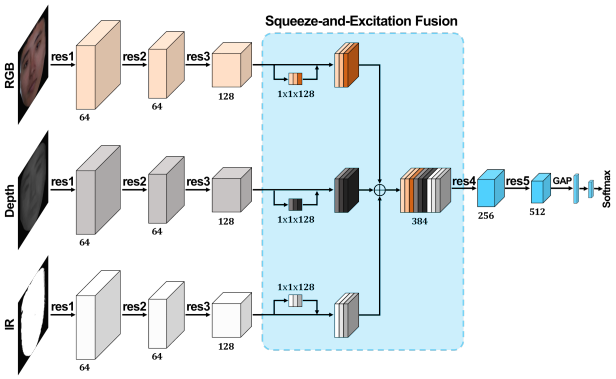

Figure 6. Diagram of the proposed fusion method. Each stream uses the ResNet-18 as the backbone, which has five convolutional blocks (*i.e.*, res1, res2, res3, res4, res5). The res1, res2, and res3 blocks are proprietary to extract the features of each modal data (*i.e.*, RGB, Depth, IR). Then, these features from different modalities are fused via the squeeze and excitation fusion module. After that, the res4 and res5 block are shared to learn more discriminatory features from the fused one. GAP means the global average pooling.

res1, res2 and res3 blocks are proprietary for each stream to extract the features of different modalities. After that, these features are fused via the squeeze and excitation fusion module. This module newly adds a branch for each modal and the branch is composed of one global average pooling layer and two consecutive fully connected layers. The squeeze and excitation fusion module performs modal-dependent feature re-weighting to select the more informative channel features while suppress less useful ones for each modal, and then concatenate these re-weighted features to the fused feature. In this way, we can make full use of the characteristics between different modalities via re-weighting their features.

5. Experiments

In this section, we describe implementation details. Then, the effectiveness of the proposed fusion method is evaluated. Next, we conduct experiments to analyze the CASIA-SURF dataset in terms of modalities and number of subjects. Finally, the generalization capability of the CASIA-SURF dataset is evaluated on standard face anti-spoofing benchmarks.

5.1. Implementation details

We resize the cropped face region to the size 112×112, and use random flipping, rotation, resizing, cropping and color distortion for data augmentation. For the CASIA-SURF dataset, all models are trained for $2,000$ iterations with an initial learning rate of 0.1, and decreased by a factor of 10 after $1,000$ and $1,500$ iterations. All models are optimized by the Stochastic Gradient Descent (SGD) algorithm on 2 TITAN X (Maxwell) GPU with a mini-batch 256. Weight decay and momentum are set to 0.0005 and 0.9, respectively.

5.2. Model Analysis

We carry out an ablation experiment on the CASIA-SURF dataset to analyze our proposed fusion method. For evaluation, we use the same settings except for the fusion way to examine how the proposed method affects final performance. From the results listed in Table 3, it can be observed that the proposed fusion method achieves TPR=96.7%, 81.8%, 56.8% @FPR=10^{-2}, 10^{-3}, 10^{-4}, respectively, which are 7.6%, 48.2% and 39.0% higher than the halfway fusion method, especially at FPR=$10^{-3}, 10^{-4}$. Besides, the APCER, NPCER and ACER are also improved from 5.6%, 3.8% and 4.7% to 3.8%, 1.0% and 2.4%, respectively. Compared with Halfway fusion method, we can verify the effectiveness of the proposed squeeze and excitation fusion method from Table 3.

5.3. Dataset Analysis

The proposed CASIA-SURF dataset has three modalities with $1,000$ subjects. In this subsection, we analyze their complementarity of importance of multi-modal and the complementarity when training with a large number of subjects.

The Effects of No. Modalities. As shown in Table 4, only using the prevailing RGB data, the results are TPR=49.3%, 16.6%, 6.8% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 8.0% (APCER), 14.5% (NPCER) and 11.3% (ACER), respectively. In contrast, simply using the IR data, the results can be improved to TPR=65.3%, 26.5%, 10.9% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 1.2% (NPCER) and 8.1% (ACER), respectively. Notably, from the numbers, we can see that the APCER of the IR data increases by a large margin, from 8.0% to 15.0%. Among these three modalities, the Depth data achieves the best performance, *i.e.*, TPR=88.3%, 27.2%, 14.1% @FPR=10^{-2}, 10^{-3}, 10^{-4}, and 5.0% (ACER), respectively. By fusing the data of arbitrary two modalities or all the three ones, we demonstrate the increasing performance. Specifically, the best results are achieved by fusing all the three modalities, improving the best results of single modality from TPR=88.3%, 27.2%, 14.1% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 5.1% (APCER), 1.2% (NPCER) and 5.0% (ACER) to TPR=96.7%, 81.8%, 56.8% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 3.8% (APCER), 1.0% (NPCER) and 2.4% (ACER), respectively. To sum up, the complementarity among different modalities will be learned. And the more numbers of modalities fused, the better performance will be.

The Effects of No. subjects. As indicated in [40], there is a logarithmic relation between the amount of training data and the performance of deep neural network methods. To

Method	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
Halfway fusion	89.1	33.6	17.8	5.6	3.8	4.7
Proposed fusion	**96.7**	**81.8**	**56.8**	**3.8**	**1.0**	**2.4**

Table 3. Effectiveness of the proposed fusion method. All models are trained in the CASIA-SURF training set and tested in the testing set.

Modal	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
RGB	49.3	16.6	6.8	8.0	14.5	11.3
Depth	88.3	27.2	14.1	5.1	4.8	5.0
IR	65.3	26.5	10.9	15.0	1.2	8.1
RGB&Depth	86.1	49.5	10.6	4.3	5.6	5.0
RGB&IR	79.1	50.9	26.1	14.4	1.6	8.0
Depth&IR	89.7	71.4	24.3	**1.5**	8.4	4.9
RGB&Depth&IR	**96.7**	**81.8**	**56.8**	3.8	**1.0**	**2.4**

Table 4. Effect on the number of modalities. All models are trained in the CASIA-SURF training set and tested on the testing set.

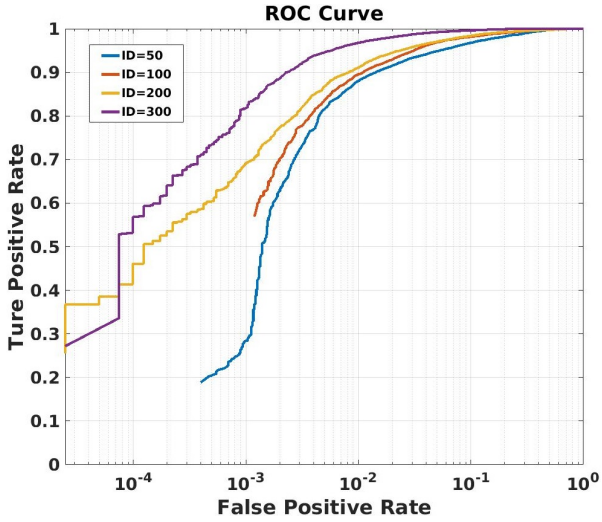

Figure 7. ROC curves of different training set size in the CASIA-SURF dataset.

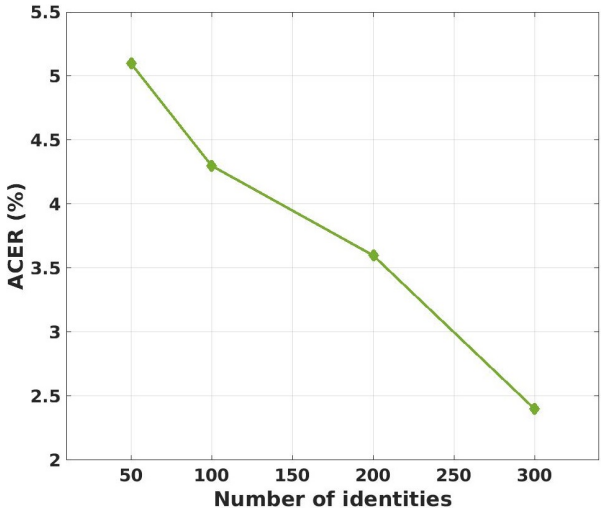

Figure 8. Performance of different training set size in the CASIA-SURF dataset.

understand the impacts of having a larger amount of training data, we show how the performance grows as training data increases in our benchmark. For this purpose, we train our baselines with different sized subsets of subjects randomly sampled from the training set. This is, we randomly select 50, 100 and 200 from 300 subjects for training. Fig.7 shows the ROC curves under different number of subjects used. We can see that when FPR is between 0 to 10^{-4}, the TPR is better when more subjects are used for training. Specially, when FAR=10^{-2}, the best TPR of 300 subjects is higher about 7% than the second best TPR result (ID=200), which demonstrates the more data used, the better performance will be. Meanwhile, in Fig. 8, we also give the performances of APCER under varied subjects used in training stage. The performance of ACER (average value of the fake and real error rates) is getting better when more subjects used. That is reasonable and rational.

5.4. Generalization Capability

In this subsection, we evaluate the generalization capability of the proposed dataset on the SiW [31] and CASIA-MFSD [49] datasets. The CASIA-SURF dataset contains not only RGB images, but also the corresponding Depth information, which is indeed beneficial to the Depth supervised face anti-spoofing methods [31, 44]. Thus, we adopt FAS-TD-SF [44] as our baseline to conduct experiments.

SiW dataset. Two state-of-the-art methods (FAS-BAS [31] and FAS-TD-SF [44]) on the SiW dataset are selected for comparison. We use the RGB and Depth images from the proposed CASIA-SURF dataset to pre-train the FAS-TD-

SF CNN model, and then fine-tune it in the SiW dataset. Table 5 shows the comparison of these three methods. FAS-TD-SF generally achieves better performance than FAS-BAS, while our pre-trained FAS-TD-SF in CASIA-SURF (FAS-TD-SF-CASIA-SURF) can further improve the performance of PAD on both protocols[1] 1, 2 and 3. Concretely, the performance of ACER is superior about 0.25%, 0.14% and 1.38% when using the proposed CASIA-SURF dataset in Protocol 1, 2, and 3, respectively. The improvement indicates that pre-training in the CASIA-SURF dataset supports the generalization on data containing variabilities in terms of (1) face pose and expression, (2) replay attack mediums, and (3) cross presentation attack instruments (PAIs), such as from print attack to replay attack. Interestingly, it also demonstrates our dataset is also useful to be used for pre-trained model when replay attack mediums cross PAIs. And the improvement is very promising ($\sim 1.38\%$).

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	FAS-BAS [31]	3.58	3.58	3.58
	FAS-TD-SF [44]	1.27	0.83	1.05
	FAS-TD-SF-CASIA-SURF	1.27	0.33	**0.80**
2	FAS-BAS [31]	0.57±0.69	0.57±0.69	0.57±0.69
	FAS-TD-SF [44]	0.33±0.27	0.29±0.39	0.31±0.28
	FAS-TD-SF-CASIA-SURF	0.08±0.17	0.25±0.22	**0.17±0.16**
3	FAS-BAS [31]	8.31±3.81	8.31±3.80	8.31±3.81
	FAS-TD-SF [44]	7.70±3.88	7.76±4.09	7.73±3.99
	FAS-TD-SF-CASIA-SURF	6.27±4.36	6.43±4.42	**6.35±4.39**

Table 5. The results of the contrast experiment in three protocols of SiW [31].

CASIA-MFSD dataset. Here we do the cross-testing experiments on the CASIA-MFSD dataset to further evaluate the generalization capability of the proposed dataset. State-of-the-art-models [11, 1, 38, 46] that use Replay-Attack [6] as the training set. Second, we train the FAS-TD-SF [44] in the SiW and CASIA-SURF datasets, respectively. All the results are shown in Table 6. It reveals that the model trained in the CASIA-SURF dataset performs the best among all results, which further validates the generalization capability of our proposed dataset.

Method	Training	Testing	HTER (%)
Motion [11]	Repaly-Attack	CASIA-MFSD	47.9
LBP [11]	Repaly-Attack	CASIA-MFSD	57.6
Motion-Mag [1]	Repaly-Attack	CASIA-MFSD	47.0
Spectral cubes [38]	Repaly-Attack	CASIA-MFSD	50.0
CNN [46]	Repaly-Attack	CASIA-MFSD	45.5
FAS-TD-SF [44]	SiW	CASIA-MFSD	39.4
FAS-TD-SF [44]	**CASIA-SURF**	**CASIA-MFSD**	**37.3**

Table 6. The results of cross testing on different cross-testing protocols.

[1]For more details of these protocols, please refer to [31].

6. Discussion

As shown in Table 3 and Table 4, we can see that even traditional evaluation metrics (APCER, NPCER and ACER) got accurate results, such as APCER=3.8%, NPCER=1.0%, ACER=2.4% in the CASIA-SURF dataset. That is say that the error rate of fake samples is 3.8% and the error rate of real samples is 1.0%. It means there is 3.8 fake samples per 100 attackers will be treated as the real one. This is very unacceptable in real applications, especially for face payment and phone unlock. Table 5 also demonstrates similar performance in the SiW dataset. In order to push the state of the arts, new lager-scale dataset and evaluation metric would be presented. Fortunately, inspired by evaluation metric widely used in face recognitoin [30], the ROC curve is widely used in academic and industry. We believe that the ROC curve is more suitable as the evaluation metric for face anti-spoofing.

As shown in Table 3 and Table 4, even the value of ACER is very promising, the TPR at different values of FPR is dramatic changing. And it is still far from the standard of practical applications that is because when FPR=10^{-4}, the TPR is only equal to 56.8%. Similar to the evaluate of a face recognition algorithm, only the TPR when FAR of 10^{-4} or 10^{-5} is meaningful for real applications [20]. Therefore, in order to decrease the gap between technology development and practical applications, the ROC curve is more suitable as the evaluation metric for face anti-spoofing. In sum up, this paper not only presents a large-scale multi-modal face anti-spoofing dataset, but also introduces new evaluation metric, which would attract more researchers and institutions to develop novel algorithms for real scenes.

7. Conclusion

In this paper, we present and release a large-scale multi-modal face anti-spoofing dataset, which is the largest one in the public datasets in terms of diversity subjects, data scale, and data modalities. We believe this dataset will push the state of the art for face anti-spoofing. Owing to the large-scale learning, we find the traditional evaluation metrics (*i.e.*, APCER, NPECR and ACER) would be not suitable in our experiments, and then introduce the ROC curve as the evaluation metric, which is more appropriate for the large-scale face anti-spoofing dataset. Besides, we present a new multi-modal fusion method, which performs modal-dependent feature re-weighting to select the more informative channel features while suppressing less useful ones for each modal. Extensive experiments have beed conducted on the CASIA-SURF dataset to verify its significance and generalization capability.

8. Acknowledgements

This work has been partially supported by Science and Technology Development Fund of Macau (Grant No. 0025/2018/A1), by the Chinese National Natural Science Foundation Projects #61502491, #61876179, by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya. We gratefully acknowledge Surfing Technology Beijing co., Ltd (www.surfing.ai) to capture and provide us this high quality dataset for this research, and also acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

References

[1] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh. Computationally efficient face spoofing detection with motion magnification. In *CVPR*, pages 105–110, 2013. 3, 8

[2] S. Bhattacharjee, A. Mohammadi, and S. Marcel. Spoofing deep face recognition with custom silicone masks. 2018. 1, 2

[3] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *TIFS*, 11(8):1818–1830, 2016. 1, 3

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing using speeded-up robust features and fisher vector encoding. *SPL*, 24(2):141–145, 2017. 1, 3

[5] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FG*, pages 612–618, 2017. 2

[6] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012. 2, 8

[7] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012. 3

[8] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel. Face recognition systems under spoofing attacks. In *Face Recognition Across the Imaging Spectrum*, pages 165–194. Springer, 2016. 2

[9] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. The replay-mobile face presentation-attack database. In *BIOSIG*, pages 1–7, 2016. 2

[10] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *ICB*, pages 1–8, 2013. 3

[11] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *ICB*, pages 1–8, 2013. 8

[12] M. De Marsico, M. Nappi, D. Riccio, and J.-L. Dugelay. Moving face spoofing detection via 3d projective invariants. In *ICB*, pages 73–78, 2012. 3

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 3

[14] N. Erdogmus and S. Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *BTAS*, pages 1–6, 2014. 2

[15] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *JVCIR*, 38:451–460, 2016. 3

[16] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 4

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[18] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 5

[19] A. Jourabloo, Y. Liu, and X. Liu. Face de-spoofing: Anti-spoofing via noise modeling. *arXiv preprint arXiv:1807.09968*, 2018. 1, 2, 3

[20] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882, 2016. 8

[21] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee. Face liveness detection using variable focusing. In *ICB*, pages 1–6, 2013. 3

[22] Y. Kim, J. Na, S. Yoon, and J. Yi. Masked fake face detection using radiance measurements. *JOSA A*, 26(4):760–766, 2009. 2

[23] D. E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 10:1755–1758, 2009. 4

[24] K. Kollreider, H. Fronthaler, and J. Bigun. Verifying liveness by multiple experts in face biometrics. In *CVPRW*, pages 1–6, 2008. 3

[25] J. Komulainen, A. Hadid, and M. Pietikainen. Context based face anti-spoofing. In *BTAS*, pages 1–8, 2013. 3

[26] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel. Complementary countermeasures for detecting scenic face spoofing attacks. In *ICB*, pages 1–7, 2013. 3

[27] N. Kose and J.-L. Dugelay. Countermeasure for the protection of face recognition systems against mask attacks. In *FG*, pages 1–6, 2013. 2

[28] J. Li, Y. Wang, T. Tan, and A. K. Jain. Live face detection based on the analysis of fourier spectra. *Biometric Technology for Human Identification*, 5404:296–304, 2004. 3

[29] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *IPTA*, pages 1–6, 2016. 3

[30] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 2, 8

[31] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398, 2018. 1, 2, 3, 5, 7, 8

[32] J. Maatta, A. Hadid, and M. Pietikainen. Face spoofing detection from single images using texture and local shape analysis. *IET biometrics*, 1(1):3–10, 2012. 3

[33] A. Mohammadi, S. Bhattacharjee, and S. Marcel. Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. *IET Biometrics*, 7(1):15–26, 2017. 1

[34] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24(7):971–987, 2002. 3

[35] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, pages 1–8, 2007. 3

[36] G. Pan, L. Sun, Z. Wu, and Y. Wang. Monocular camera-based face liveness detection by combining eyeblink and scene context. *Telecommunication Systems*, pages 215–225, 2011. 3

[37] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *TIFS*, 11(10):2268–2283, 2016. 3

[38] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *TIP*, 24(12):4726–4740, 2015. 8

[39] W. R. Schwartz, A. Rocha, and H. Pedrini. Face spoofing detection through partial least squares and low-level descriptors. In *IJCB*, pages 1–8, 2011. 3

[40] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852, 2017. 6

[41] R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, S. Ricerche, and F. Roli. Fusion of multiple clues for photo-attack detection in face recognition systems. In *IJCB*, pages 1–6, 2011. 3

[42] L. Wang, X. Ding, and C. Fang. Face live detection method based on physiological motion analysis. *Tsinghua Science & Technology*, 14(6):685–690, 2009. 3

[43] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection using 3d structure recovered from a single camera. In *ICB*, pages 1–6, 2013. 3

[44] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, and Z. Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *arXiv preprint arXiv:1811.05118*, 2018. 7, 8

[45] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *TIFS*, 10(4):746–761, 2015. 1, 2

[46] J. Yang, Z. Lei, and S. Z. Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014. 3, 8

[47] J. Yang, Z. Lei, S. Liao, and S. Z. Li. Face liveness detection with component dependent descriptor. In *ICB*, volume 1, page 2, 2013. 3

[48] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1

[49] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *ICB*, pages 26–31, 2012. 2, 7